

A norm on homology of surfaces and counting simple geodesics

By GREG MCSHANE and IGOR RIVIN*

Introduction

Let S be a hyperbolic surface of finite volume, and define $N_S(L)$ to be the number of simple (that is, without self-intersections) closed geodesics on S whose length does not exceed L . This quantity arises naturally in a number of contexts, and hence it is useful to obtain some estimates on the order of growth of $N_S(L)$ as L grows large. Although a number of people have worked on this question, the answers have not, on the whole, been completely satisfactory. For general cusped S Birman and Series [2] have shown that the $N_S(L)$ grows at most polynomially as a function of L . They bound the degree of the polynomial by a function of the genus, and note that the bound is very far from optimal. For the simplest hyperbolic surface – the once punctured torus, the best result to date has been that of Beardon, Lehner, and Sheingorn [1], who showed that there $N_S(L)$ grows at least linearly, and at most quadratically in L . The case of the punctured torus is interesting for reasons other than its relative simplicity – simple geodesics on the punctured torus have deep connections with diophantine approximation, see for example [5].

The purpose of this note is to prove the following estimate:

THEOREM 0.1. *Let S be a punctured torus equipped with a complete hyperbolic metric of finite volume. Then, $\lim_{L \rightarrow \infty} N_S(L) = c_S(L^2) + O(L \log L)$, where c_S depends on the hyperbolic metric.*

The proof of Theorem 0.1 is based on some preliminaries from hyperbolic geometry (Section 1), topology of surfaces (Section 2), the general metric theory of groups acting on spaces (section 3). The final ingredient is a norm on the first homology of surfaces coming from the study of curve systems (Section 4, and subsequent).

*Igor Rivin has been supported in part by the Institute for Advanced Study under the NSF Grant #DMS-9304580. Igor Rivin would like to thank Caltech for its hospitality during the preparation of this paper

Note. An estimate of type 0.1 for the modular torus (but stated in the language of Markoff spectra) is obtained by Don Zagier in ???. Zagier's methods are computational. Geoff Mess told the authors that he has independently discovered the norm on homology used here. Mess' result has not been published.

1. Simple geodesics do not enter cusps

The result of this section goes back to Poincaré, while the sharp version for the torus stated below in Theorem 1.1 is proved in Greg McShane's 1991 Warwick thesis [6]. First, a definition:

Definition. A *cuspidal region* is a neighborhood of the cusp of S , bounded by a horocycle.

THEOREM 1.1. *Let $\epsilon > 0$. Any punctured torus has a cuspidal region with bounding curve of length $4 - \epsilon$, and this bound is optimal. No simple closed geodesic intersects a cuspidal region with boundary curve of length $4 - \epsilon$.*

For current purposes it will be sufficient to prove the following very easy theorem, weaker than Theorem 1.1 in the case of a punctured torus.

THEOREM 1.2. *Any cusped hyperbolic surface S has a cuspidal region with bounding curve of length 2. No simple closed geodesic intersects a cuspidal region with boundary curve of length 2.*

Proof. In the upper half-plane model of H^2 , consider a fundamental domain of S , arranged in such a way that the parabolic element preserving the cusp in question is $\lambda : z \rightarrow z + 1$. A simple closed geodesic g of S lifts to a geodesic in H^2 , which is represented by a semicircle \tilde{g} in the half-space model. Since g is simple, it follows that $\tilde{g} \cap \lambda(\tilde{g}) = \emptyset$. Thus, the radius of the semicircle representing \tilde{g} is smaller than $\frac{1}{2}$. The same goes for any non-vertical boundary component of the fundamental domain of S , and thus the horocyclic arc joining (for example) $\frac{1}{2}i$ and $1 + \frac{1}{2}i$ is entirely contained in the fundamental domain of S , and meets no lift of any closed geodesic of S . This horocyclic arc (which projects to a closed horocycle in S has length 2. \square

2. Primitive elements in the fundamental group of the punctured torus

The reader should recall that the fundamental group $\pi_1(T)$ of the punctured torus is the free group on two generators $F_2 = \langle s, t \rangle$, where s and t are the standard generators.

Definition. An element $\gamma \in F_2$ is called primitive, if there exists an automorphism ϕ of F_2 , such that $\psi(\gamma) = s$. If $\phi(\delta) = t$, then γ and δ are called *associated primitives*.

FACT 2.1. *The outer automorphism group of F_2 is isomorphic to the mapping class group of the punctured torus.*

This was apparently first observed by Max Dehn, though Fact 2.1 has now passed into the folklore. One argument proceeds roughly as follows:

The outer automorphism group is generated by the so-called Nielsen transformation, which either permute the basis x, y , or transform it into xy, y , or xy^{-1}, y . In the case of a punctured torus these transformations can be topologically realized by Dehn twists.

FACT 2.2. *Let $\psi : F_2 \rightarrow \mathbb{Z}^2$ be the canonical abelianizing homomorphism. Then, if γ_1 and γ_2 are two primitive elements in F_2 such that $\psi\gamma_1 = \psi\gamma_2$, then γ_1 and γ_2 are conjugate.*

Fact 2.2 follows easily from the work of Nielsen. For the proof see [8].

The following result, and its proof seem to go back to Poincaré.

FACT 2.3. *Let S be a hyperbolic surface of finite volume, and let γ be a non-trivial simple closed curve, whose corresponding covering transformation Γ is hyperbolic. Then there is a unique geodesic freely homotopic to γ ; this geodesic is simple.*

Proof. The existence of a unique geodesic $\tilde{\gamma}$ freely homotopic to γ follows by a completely standard straightening argument: since Γ is hyperbolic, it has two fixed points on the circle at infinity of \mathbf{H}^2 , as do all of its translates. The sought after geodesic $\tilde{\gamma}$ is the unique geodesic in \mathbf{H}^2 joining those two fixed points. To show that $\tilde{\gamma}$ is simple is also not hard. Indeed, the simplicity of γ implies that for any covering transformation β , the fixed points of $\beta\Gamma\beta^{-1}$ do not separate those of Γ in the cyclic order at infinity (otherwise the corresponding lifts of γ would have intersected.) However, this immediately implies that the straightened curves do not intersect, by looking at it in the projective model of \mathbf{H}^2 . \square

LEMMA 2.4. *No simple geodesic on the punctured torus separates.*

Proof. Suppose there was a separating simple geodesic γ . Let it cut the torus into two components T_1 and T_2 , and assume that T_1 is the component which does not contain the cusp. Double T_1 along its boundary to obtain a compact oriented hyperbolic surface $S = 2T_1$. The area of S is at least 4π , and thus the area of T_1 is at least 2π . But the area of the whole punctured torus is exactly 2π , and thus γ could not have been separating. \square

Note. The above argument can be easily adapted to show that there are no simple closed geodesics on the thrice-punctured sphere.

The following lemma is well known:

LEMMA 2.5. *For any pair of non-separating simple closed curves γ_1 and γ_2 on the punctured torus T , there exists a homeomorphism of T taking γ_1 to γ_2 .*

Proof. This is immediate from the classification of surfaces. \square

Fact 2.2 and Lemmas 2.4 and 2.5 combine to show that is that there is a one-to-one correspondence between simple geodesics and primitive homology classes on the punctured torus (a primitive homology class is a class (m, n) , such that m and n are relatively prime), and it is geometrically obvious that this geodesic is really none other than the familiar (m, n) torus knot, or the (m, n) geodesic on the flat torus (without the puncture). The precise result is given by the following construction of Osborne and Zieschang [8]:

Let m and n be a pair of relatively prime non-negative integers. Define a function $f_{m,n} : \mathbb{Z} \rightarrow \{1, 2\}$ as follows: $f_{m,n}(k) = f_{m,n}(k')$, if $k = k'$ modulo $m + n$; for k in $\{1, \dots, m\}$ let $f_{m,n}(k) = 1$, and for k in $\{m + 1, \dots, m + n\}$ let $f_{m,n}(k) = 2$.

Now let

$$W_{m,n}(x_1, x_2) = \prod_{i=0}^{m+n-1} x_{f_{m,n}(1+im)}.$$

If $m < 0$, let $W_{m,n}(x_1, x_2) = W_{-m,n}(x_1^{-1}, x_2)$, and if $n < 0$, let $W_{m,n}(x_1, x_2) = W_{m,-n}(x_1, x_2^{-1})$.

The main theorem of [8] is the following:

THEOREM 2.6. *If x_1 and x_2 are associated primitives in $F_2 = \langle s, t \rangle$, and if m and n are relatively prime, then $W_{m,n}(x_1, x_2)$ is a primitive. furthermore, if $mq - np = 1$ then $W_{m,n}$ and $W_{p,q}$ are associated primitives. In particular, up to conjugation the primitives of $F(s, t)$ are $\{W_{m,n}(s, t) | m, n \in \mathbb{Z}, (m, n) = 1\}$.*

It can be seen that the words $W_{m,n}(s, t)$ are all cyclically reduced, hence of minimal word length in their conjugacy class. Since the word length of $W_{m,n}$ is equal to $m + n$ the following combinatorial version of Theorem 0.1 holds:

THEOREM 2.7. *The number of conjugacy classes of primitive elements in $F(s, t)$ with reduced length not exceeding L is asymptotic to L^2 .*

Proof. The number conjugacy classes in question equals

$$f(L) = |\{(m, n) | m + n \leq L, (m, n) = 1\}|.$$

It follows from elementary number theory that $f(L)$ is asymptotically equal to $L^2/(2\zeta(2))$. \square

Remark 2.8. For free groups of higher rank the analogue of Theorem 2.7 fails spectacularly, in that the number of primitive elements of reduced length not exceeding L grows exponentially, as demonstrated by the following construction due to Casson [4]. Let F_n be the free group generated by x_1, \dots, x_n . Now let y_i , $1 \leq i \leq n$ be defined as follows: $y_i = x_i$ when $i < n$, and $y_n = x_n \prod x_{i_k}^{j_k}$, where $i_k < n$ for all k , $i_k \neq i_{k+1}$. Clearly, $\{y_i\}$ is a generating set for F_n , and also each y_i is a cyclically reduced word. It is also clear that the number of possibilities for y_n of length not exceeding L grows exponentially in L , for $n > 2$.

3. Geometry of group actions

First, some definitions and a theorem, all directly from [3].

Definition. A geometry is a metric space in which each bounded set has compact closure. A group G acts *geometrically* on a space X if X is a geometry, and there is a homomorphism ϕ (usually suppressed) from G into the isometry group of X such that the G -action is properly discontinuous and cocompact (*properly discontinuous* means that for each compact set K in X , the set

$$\{g \in G \mid \emptyset \neq K \cap gK\}$$

is finite), and *cocompact* means that the orbit space X/G is compact).

Definition. An intrinsic metric on a path space X is one where the distance between two points is the infimum of path lengths between those points.

Definition. A relation $R : X \rightarrow Y$ between spaces X and Y is said to be quasi-Lipschitz if R is everywhere defined, and there exist positive numbers K and L , such that for each $A \subset X$,

$$\text{diam } R(A) \leq K \text{ diam } A + L.$$

Definition. Relations $R : X \rightarrow Y$ and $S : Y \rightarrow X$ are *quasi-inverses* if they are everywhere defined and there exists a constant $M > 0$, such that $d(S \circ R, \text{id}_X) < M$ and $d(R \circ S, \text{id}_Y) < M$.

Definition. A relation $R : X \rightarrow Y$ is a *quasi-Lipschitz equivalence* if there is a quasi-inverse $S : Y \rightarrow X$, such that both R and S are quasi-Lipschitz.

Finally, a theorem:

THEOREM 3.1. *If a group acts geometrically on two geometries X and Y with intrinsic metrics, then X and Y are quasi-Lipschitz equivalent.*

Having dispensed with the generalities, an observation:

OBSERVATION 3.2. *If a finitely presented group G is equipped with the word metric, then G is a geometry, on which G acts geometrically by multiplication from the right.*

Now, restrict to the case of current interest, where $G = F_2$ acts on the hyperbolic plane \mathbf{H}^2 with quotient the punctured torus. If this action was geometric, in the sense of the above definitions, then Theorem 0.1 would follow from Theorem 2.7, Theorem 3.1, and the Observation 3.2. However, the action is not geometric, since the hyperbolic punctured torus is not compact. It is not hard to construct the right geometry.

Consider the tessellation of \mathbf{H}^2 by fundamental domains for the given F_2 action, and truncate each fundamental domain by a closed horocycle of length 2 (as is possible by Theorem 1.2). The geometry X will be the subset of \mathbf{H}^2 equal to the union of all these truncated fundamental domains, and equipped with path metric. Clearly X is a geometry, on which F_2 acts geometrically, and so X is quasi-isometric (quasi-Lipschitz equivalent) to F_2 equipped with its word metric.

While the general minimal paths on X may be quite far from hyperbolic geodesics, by Theorem 1.2, the minimal paths corresponding to lifts of simple closed geodesics are unaffected by the truncation, and so their length are within a constant factor of the length of the corresponding reduced words in the word metric.

4. Geometry of multicurves

Let S be a hyperbolic surface of finite volume with at most one cusp. We define a *multicurve* m on S to be a map from a (not necessarily connected) 1-manifold M to S . We define the *length* of m to be the sum of the lengths of the images of components of M . We say that a multicurve m is *embedded* if the image of m is the union of simple closed curves $\gamma_1, \dots, \gamma_k$ on S . Note that the map m may cover some of the components γ_i multiple times, so this does not coincide with the usual meaning of embedding. A multicurve defines a singular chain, which, in turn, defines a homology class in $H_1(S, \mathbb{Z})$.

The first observation is the following:

THEOREM 4.1. *Let $h \in H_1(S, \mathbb{Z})$ be a non-trivial homology class. There exists a multicurve m representing h of minimal length, and m is embedded, with all components geodesic.*

Proof. First, we show the existence. If S is a compact surface, this follows by a standard Arzela-Ascoli argument. If S has cusps, let l_h be the infimum of the lengths of multicurves representing h ; clearly $l_h > 0$. Let m_1, \dots, m_l, \dots be multicurves whose lengths approach l_h . In order to apply the Arzela-Ascoli theorem, it is enough to show that the diameters of the images of the m_i are uniformly bounded. Since the lengths of m_i approach l_h , it follows that the diameters of all the components of m_i , for i sufficiently great, are uniformly bounded by $2l_h$. On the other hand, we can assume that no component of m_i , for i sufficiently large, can be contained entirely in a horodisk of area 2 surrounding each cusp, since such a component is homologically trivial, and thus we can delete such a component with no change to homology, and replace m_i by the resulting multicurve m'_i . Hence, we can assume that all of the m_i are contained in a compact subset of S , and thus the existence follows.

Now let m be a multicurve of minimal length representing h . It is clear that each component of m is geodesic. Suppose that m is not embedded, thus two components c_1 and c_2 of m intersect. We can assume that c_1 and c_2 are not multiply covered (if they are, we can split off one circle off each), and since they are geodesic, the intersection is transverse. Let O be an intersection of c_1 and c_2 , and let A_1OB_1 and A_2OB_2 be small directed segments of c_1 and c_2 , respectively, surrounding O . By cutting and pasting, we can replace these by $s_1 = A_1OB_2$ and $s_2 = A_2OB_1$, without changing the homology, and then by smoothing s_1 and s_2 at O , we obtain a shorter multicurve than m representing h , thus arriving at a contradiction. \square

COROLLARY 1. *Let T be a punctured torus equipped with a hyperbolic structure. Then, the shortest multicurve representing a non-trivial homology class h is a simple closed geodesic if h is a primitive homology class (that is, not a multiple of another class), and a multiply covered geodesic otherwise. In addition, the shortest multicurve representing h is unique.*

Proof. Since there is exactly one hyperbolic geodesic in any homotopy class, it is enough to observe that any two non-homotopic curves on a punctured torus intersect. Theorem 4.1 then implies that the shortest multicurve representing h has one component, which is multiply covered if h is not primitive (if $h = (m, n)$, with $d = \gcd(m, n)$, then the shortest multicurve representing h is covered d times). Since there is at most one simple closed geodesic in a homology class (see, eg section 2), the uniqueness follows. \square

5. A norm on homology of the punctured torus

First, let us define a valuation ℓ on $H_1(T, \mathbb{Z})$, where $\ell(h)$ is defined to be the length of the shortest multicurve representing h . The valuation of the

trivial homology class is defined to be 0. Corollary 4.1 implies that

$$(5.1) \quad \ell(nh) = n\ell(h),$$

and that

$$(5.2) \quad \ell(h + g) \leq \ell(h) + \ell(g),$$

where the inequality is strict if h and g are not both multiples of the same homology class. The latter inequality holds, because the union of the shortest multicurves corresponding to h and g is not embedded, and hence, the shortest multicurve corresponding to $h + g$ is shorter than the union. It follows that ℓ can be extended to the rational homology $H_1(T, \mathbb{Q})$, by linearity (equation 5.1), and further, to $H_1(T, \mathbf{R})$ by continuity, which follows from equations 5.1 and 5.2. Since $\ell(0) = 0$, and by equations 5.1 and 5.2, ℓ is a pseudo-norm on the two-dimensional vector space $H_1(T, \mathbf{R})$. By the results of sections 1, 2 and 3, ℓ is actually a norm, since the results of those sections imply that $0 < c_1 < \ell(h)/\|h\|_1 < c_2 < \infty$, where $\|h\|_1 = |m| + |n|$ for $h = (m, n)$. It follows that the unit ball of the norm ℓ is a compact, convex figure B_ℓ in the plane, and the number of simple geodesics of length not exceeding L on the torus T is equal to the number of primitive lattice points in LB_ℓ . Thus, Theorem 0.1 follows.

6. Further investigations.

The geometry of the unit ball of the norm ℓ is very interesting; the authors will discuss it in a future paper. In particular, the study of the unit ball can be used to show that the error term in Theorem 0.1 is essentially sharp. See [7].

The methods of this paper can be extended without much difficulty to define a norm on homology of surfaces of higher genus, by extending the length of shortest multicurve valuation, and this can be used to count minimal multicurves. It seems non-trivial to use this to count simple geodesics on surfaces of higher genus. Using the theory of train tracks and the methods of this note it can be shown (McShane and Rivin, in preparation) that the number of geodesics of length not exceeding L on a closed surface of genus g with c cusps has order of growth $L^{6g-6+2c}$.

REFERENCES

- [1] A. F. Beardon, J. Lehner, and M. Sheingorn. *Closed geodesics on a Riemann surface with applications to the Markov spectrum*, Transactions of the American Mathematical Society **295**(2), 1986.
- [2] Joan Birman and Caroline Series. *Geodesics with bounded intersection are sparse*, Topology **24**(2), 1985.
- [3] James Cannon. *Geometric groups*, appendix to *The theory of negatively curved spaces and groups*, Ergodic theory, dynamical systems, and hyperbolic geometry, collection edited by Tim Bedford, Michael Keane and Caroline Series, Oxford University Press, 1991
- [4] Andrew Casson. *Personal communication*, 1994.
- [5] Andrew Haas. *Diophantine approximation on hyperbolic Riemann surfaces*, Acta Mathematica, 1986.
- [6] Gregory McShane. *A remarkable identity for lengths of curves*, Doctoral dissertation, University of Warwick, Coventry, 1991.
- [7] Gregory McShane and Igor Rivin. *Simple curves on hyperbolic tori*, preprint, 1994.
- [8] R. P. Osborne and H. Zieschang. *Primitives in the free group on two generators*, Inventiones Mathematicae, **63**, 1981.
- [9] Don Zagier. *On the number of Markoff numbers below a given bound*, Mathematics of Computation **39**, 1982.